chapter

1

人文情報学と歴史学

後藤 真（国立歴史民俗博物館）

1. はじめに－人文情報学とは何か

人文情報学とは、そもそも「人文学に情報学の技法や技術を応用する学問」として定義されます。人文学のある課題に対して、情報学のさまざまな技術を適用することで、人文学がこれまでに見つけてこなかった知見を新たに発見できる、もしくは人間の行う作業を省力化することで人間にしかできない研究の時間などをなるべく確保し、新たな知見への可能性を高めることを目指す学問です。なお、少なくとも本稿執筆現在ではあまり一般的ではないですが、情報学の研究開発に人文学の成果を用いるということも、この学問のスコープに入っています。例えば、AI のさまざまな分析に人文学的な倫理観などの思考を入れることが対象となるでしょう。または、コンピュータプログラムの技術に対して、人文学の技法そのものを当てはめるなどの議論も考えられます（無論、コンピュータプログラムの多くは人間の思考を抽象化したものですが、実際にこれらの新規開発において具体的な人文学とのコラボレーションはありません）。したがって、多くの場合は上記の通り、人文学の成果への技術応用と考えて差し支えはないでしょう。

日本において、歴史学を正面から捉えつつ、歴史情報学の研究を行っている組織は、現時点では国立歴史民俗博物館のみ、とも（ややいいすぎかと思い

ますが）いえます。しかし、無論、歴史情報学を包含する人文情報学全体や、情報学の文脈において、歴史学と情報学に関わる研究を進めている研究組織や学会はあまたあり、そこで重要な研究が多く生まれてきているのは事実です。さらに歴史学の文脈からも、特にデータベースの構築に関わった重要な研究が生まれてきつつあります。そして、デジタルアーカイブという大きなうねりの中でも、歴史資料の高度なデジタル化が進められつつあるのが現状です。本章では、このような歴史学に関わる人文情報学の研究状況を概観し、さまざまな研究の情報を得ることを目指すことにしましょう。なお、デジタルアーカイブに関わる状況については第９章で述べるため、この章では触れません。

2. 日本における歴史情報学・人文情報学の研究状況

まずは、一般的な人文情報学の傾向について述べるとともに、その研究の中心となる学会について簡単に述べておくことにします。人文情報学の研究傾向としては、大きくは発見系･解析系･可視化系の３つの系列があるといってよいと考えます。

発見系は人文系の大規模な情報群・データ群の中から必要なものをいかに効率的に発見することができるか、もしくは研究の目的に沿って見つけることができるかを研究する系統にあたります。いわば人文系研究の「スタート」の部分をフォローすることを中心とした研究の系統であるといってよいでしょう。

解析系は、テキストや絵画などをデータ化し、それらのデータについて、コンピュータプログラムを用いて分析し、人間の手法と異なるアプローチを行う系統です。解析系は人文学が行っている研究の中でも､特に分析「経過」を新たなかたちにしていくことを目指したものであるといえるでしょう。

可視化系は、研究成果を中心にいかにわかりやすく研究者以外の人びとに見せるかなどを研究するものです。特に、人文系研究の最終的な「出口」の

部分で、新たなかたちを担うものにあたります。

一人の研究者やひとつのグループが、どれかひとつの系統のみを研究することということではありません。例えば、テキストデータを使ってある特定の文章の傾向を導き出すとともに、検索結果などについて可視化を行ったり、情報発見の手法にテキスト解析の成果を取り入れたり、異なる系統の複数のアプローチを取ることは十分にあり得ます。ただし、多くの場合は、この3つのうちどれかひとつの方向を特に中心的なテーマとし、それに派生するかたちで研究をするという傾向が多いのは確かです。

発見系の研究としては、近年の重要なキーワードにLinked Dataやセマンティック Webなどがあげられるでしょうか。これ以外にも巨大なデータベースの構築手法や長期保存も、この研究の文脈に入ります。本書で触れるTEI・IIIFも比較的近い方向ですが、TEIは解析にも用いられることが多いです。

解析系の研究としては、テキストの計量解析がもはや「古典」といってよいものになっています。本章の後半でも触れますが、KH Coderのような便利なツールも生まれていますし、それらのツールを使い、比較的容易に研究を進めることができるようになっています。また、画像解析などもディープラーニングの進展によって進んできています。近年だとくずし字の解析やOCRなどへの活用も、この系統にあたるでしょうか。

可視化系の研究としては、これまではGISのような、資料情報の空間上へのプロットなどが大きなテーマでした。これに加え3Dでの表現技法などの研究が進んでいます。近年はデバイスの発達によって、VRやARのような研究も大きく盛り上がってきています。

本書では、コラムを除くと、発見系の研究を中心に取り上げることになります。それは、日本における歴史情報の研究はほとんどの場合、発見系だからです。

この理由はいくつかあります。ひとつには、そもそも歴史学が対象とする資料の「幅広さ」に起因します。歴史学がひとつの研究を行う際に対象とす

る資料は、非常に幅が広いのが特徴です。例えば、ある研究を行うのに日記資料のようなまとまった文章になったものと、古文書のような手紙類、法律関係などの公文書を同時に、それも研究上同じ重さを持って扱うことは、極めて普通の行為です。その場合、大量にある資料の中からそれらの情報を発見することが重要であり、その発見行為をサポートすることが、歴史学研究に重要な貢献をもたらすといえる部分があります。また、この「幅広さ」は解析系の研究の大きな障壁になっているのも事実です。計量解析を行う際に、比較的似た性質の資料を扱うことができれば、前提となる知識が同じになるため、解析の意味を読み取りやすいのですが、性質の大きく異なる資料を対象とする場合には、コンピュータによる結果を判断するのが困難になります。例えば、ある人物Aの日記と人物Bの手紙群の文体解析を行いたいとします。両者とも一定のテキスト量があったとして、その頻出語句の解析自体は重要になるかもしれません。しかし、その両者を比較しようとしても、そもそも書いている文章のあり方が違いすぎるので、成立しないということになります。両者の日記が比較できればよいのですが、往々にして人物Aは日記だけで手紙類は残っていない、人物Bは日記がない、といった事態が起こります。そのため解析研究が進みにくいという側面があります（言い換えれば、解析研究を行う際には、前提としてその資料の理解をしっかりと行っておく必要があるといえます）。

もうひとつの特徴は、そもそもデジタル化の遅れがあり、デジタル化を推進するとともに研究を実施しているという側面があります。歴史資料に関するデジタルデータは、決して多いとはいえません。特に解析に使える資料群は少なく、基本的なデータをしっかりと作るために、多くの資料を持っている研究者が努力している状況です。歴史情報に関わる研究者が結果的に資料所蔵の機関にかたよってしまっているという側面もあるでしょう。そのため、その情報基盤を発見するための研究が多いともいえます。可視化に関する研究もありますが、博物館などでの成果の公開という側面が大きくなっており、

一部のサンプルでの可視化にとどまる例が多いです。

言い換えれば、これらの課題を突破し、分析系の研究を行うことができるならば、それは特に若い研究者にとってみれば大きな可能性になるともいえるでしょう。特にテキストが持つ特殊性のような課題を乗り越えることで、新たな研究へとつなげられる部分があるともいえます。無論、海外、特に英語圏については多くの研究蓄積がありますので、それらの事例を参考にしつつ、日本の歴史資料に当てはめてみることはできるでしょう。

3. 情報を得るために人文情報に関わる学会とその傾向

次に、これらの研究に関する情報を取得したり、発表するための学会について述べてみたいと思います。特に人文情報（Digital Humanities）に関するものがその中心になります。

まず、世界レベルでは Alliance of Digital Humanities Organization（ADHO）と呼ばれる学会連合での活動が中心となっています。ADHO はヨーロッパにおける Association for Literary and Linguistic Computing（ALLC）、米国での Association for Computers and the Humanities（ACH）、カナダの Society for Digital Humanities/Société pour l'étude des médias interactifs（SDH/SEMI）と呼ばれる学会が連合して作られたものです。その後、世界各地の DH 学会の連合体となっており、日本においても JADH（Japanese Association for Digital Humanities）という団体がこの ADHO に加盟しています。ADHO は年に一度国際会議を開き、人文情報学におけるトップカンファレンスとして位置づけられています。国際的には、まずこの学会で発表するということがひとつの目標になるでしょう。2019 年の大会（2018 年に投稿したもの）では、1000 に近い投稿があるなど、大変に活発であるとともに、採択されるのも大変になりつつありますが、ここへの投稿をまずは目指しましょう。

また、その前段階として、日本国内で英語で行われる国際会議である JADH で発表するというステップを踏むことも必要です。

そして、日本国内では、この JADH のほかに情報処理学会の研究会（SIG）である人文科学とコンピュータ研究会（以下、CH 研究会と呼びます）というものがあります。CH 研究会は人文情報学における最も「老舗」の研究会のひとつであり、1999 年より年に一度国内では最も大きなシンポジウムである「人文科学とコンピュータシンポジウム（通称「じんもんこん」）」を開催するとともに、年に 3 回（2016 年度までは 4 回）の通常の研究会を行っています。この CH 研究会は対象とする分野も広く、歴史のみならず、人文情報の観点から発見・解析・可視化のすべての研究が満遍なく見られる点が特徴です。日本における人文情報学の現状を、特に技術的な文脈も含めて確認する際には、この CH 研究会の研究報告などを確認することから始めると研究史を含めて広く理解できるでしょう。とりわけ、年に一度開かれる「じんもんこん」は、日本における人文情報・歴史情報の最新の動向を見る上で、最も重要な会議といえます。また、この研究会の全体の学会となる情報処理学会の論文誌においても、人文科学とコンピュータに関わる特集が組まれるようになっています。情報処理学会の論文誌は日本国内では比重の大きなものですので、特に技術的な新規性をもととした研究で、日本語での成果発表を行いたい場合には、ここがひとつの目標になります。

日本における CH 研究会と海外の DH の研究は当然密接なつながりがあります。しかし、このふたつには若干傾向の違いも見られます。その最も大きな違いは、CH 研究会では発見系・可視化系の系統がこれまで比較的研究の数としては多い傾向があったのに対して、DH は解析系が優位だったことです。CH 研究会においては解析系、とりわけテキスト解析が決して多いわけではない状況が長く続いていました。この理由としては、CH が情報処理学会という情報工学の研究の中に位置づけられていることがあげられます。一方で DH はもともと人文系のテキスト分析などの研究者がコンピュータを応用することから始まっていたこともあり、分析系の研究が非常に多いです。とりわけ、解析ツールの開発とテキストデータの標準化などの研究が多

いという傾向を持っています。しかし、両者の交流が密接に進んだ近年ではそれらの差異も減りつつあります。CH でも解析系研究が多くなり、一大潮流となっている一方で、JADH などでも発見・可視化の研究が多く発表されるなどその間の差はなくなりつつあるのが現状です。しかし、「どのような読者が多いのか」といった傾向や、比較的古い論文を探す際には、その傾向は知っておくとよいでしょう。

CH 研究会や DH の学会以外にも、複数の学会があります。情報知識学会や、アート・ドキュメンテーション学会、デジタルアーカイブ学会などが国内学会としては代表的な存在でしょう。

情報知識学会は、特に図書館情報学に近い分野での研究が多い点が特徴で、研究成果の解析などもその対象となっています。また、リポジトリなどの研究も、情報知識学会の中では深く議論されています。この傾向は CH 研究会ではあまり見られません。情報知識学会は年に一度の大会と、同じく年に一度、企画を中心とした「情報知識フォーラム」を開催しています。さらに、情報知識学会は年に 4 冊の論文誌を発行しており、コンスタントに論文というかたちでの情報を得ることができるとともに、迅速な学会での成果公開が行われている点も特徴といえるでしょう。

アート・ドキュメンテーション学会は、特に美術館・博物館の関係者が多いのが特徴です。そのため、特にメタデータなどの議論が盛んであり、さらに現場の現状に即した研究報告も多くされています。この点では、日本でも他学会と異なる独自の研究方向があるといえるでしょう。年に一度の大会と、秋季に研究集会が行われており、学会誌としては『アート・ドキュメンテーション研究』を刊行しています。

デジタルアーカイブ学会は、デジタルアーカイブの近年の大きな隆盛を受けて作られた学会です（2017 年設立）。この学会は、裾野が広いことが特徴として指摘できますが、これまで紹介してきた学会との大きな違いは、法制度部会があるという点だと考えます。そのため、この学会では特にオープンデー

タや、デジタルアーカイブの著作権などの問題を、学会の中で広く取り扱っており、この点は他の人文情報系の学会では見られない取り組みです。年に一度の研究大会を開催し、『デジタルアーカイブ学会誌』を刊行しています。

このような学会が人的交流を持ちつつ、研究を進めています。それぞれの学会が、完全に機能特化しているわけではありません。それぞれに跨りながら研究を進めています。しかし、各学会の特徴を把握しておくことで、自身の研究に関わる情報入手や、研究発表による助言をより効果的・効率的に得ることができるでしょう。

海外の学会を含む全体のリストを本書の末尾に置くので参考にしてください。

4. どのような研究があるか－ツールとデータベース

次に、日本で開発され、特に人文情報学の学会などで発表されている、歴史情報学の研究に使用できるデジタルツール類を紹介します。あとで紹介する大規模データベースなどの事例の量に対して、デジタルツールの開発は、現時点ではそれほど多いとはいえません。しかし、知っておくことで研究の進展に有益なものもありますし、ツール開発などのためにはどのような研究をすればよいかのヒントとしても有益です。

▶ 時間情報解析ソフトウエア　HuTime（http://www.hutime.jp/）

特に日付に関してさまざまな分析を可能としているソフトウエアとその関連プロジェクトを HuTime と呼称しています。大きくは 3 つのツールに分かれます。ひとつは、時間情報をさまざまなかたちで可視化するデスクトップアプリケーションツール。もうひとつは、日本における時間の表記を変換し、コンピュータによりわかりやすい形式にする Web サービス、そして時間情報に関する RDF データを提供しているシステムになります。これらのシステムにより、和暦の表記などを西暦などに変換し、コンピュータでも使用可能にすることができる点が特徴です。このシステムについては、時間情報の

基盤データとしても活用することができるでしょう。

▶ 日本語のテキスト解析ソフトウエア　KH Coder（http://khcoder.net/）

日本語について、形態素解析などの分析を行うとともに、それらを可視化することができるツールです。KH Coder は非常に多くの可視化ツールを持っており、日本語資料の基礎的な分析を行う場合には、このツールを用いることから始めるとよいでしょう。このツールで何ができるのか、そして、解析を行うことで何が見えてくるのかなどを知ることができます。

▶ 国立国語研究所による日本語の歴史的コーパス

（https://pj.ninjal.ac.jp/corpus_center/chj/）

国立国語研究所は、日本語のさまざまなコーパスを提供しており、広く歴史研究に使えるものもあります。「日本語歴史コーパス」は、奈良時代から明治・大正に至るまでのコーパスを広く提供しています。また、「オックスフォード・NINJAL 上代語コーパス」は上記のコーパスよりさらに古く、万葉集など、奈良時代以前の和歌を中心としたコーパスデータとして構築されており、日本語の歴史的な資料を分析する前の基本的なデータとして有益です。

▶ 日本語のくずし字練習ツール「KuLa」

「みんなで翻刻（ほんこく）」（https://honkoku.org/）を作成した橋本雄太（はしもとゆうた）氏が開発した、古文書を読むために必要な「くずし字」を学ぶためのスマートフォンアプリケーションです。変体仮名と基本的な漢字について、テスト機能を使いながら学ぶことができます。このように、歴史資料を勉強するためのツール開発も重要なひとつの分野になっています。現状では実験的なものが多いですが、必要な資料を取り込み、最終的なアプリケーションにしたという点においては、KuLa はひとつのベンチマークといえるでしょう。

これらのツールは、無論、歴史研究のみに限らず、広く人文情報学の中で注目される案件です。当然、日本における歴史資料を分析する際にも有用な

ものとなっています。

次に、関連する大型のデータベースの紹介をします。これらは、まさに発見系研究を進める中心的な拠点であり、データ発見の技法には、人文情報学の多くの知見が活かされています。

▶ 東京大学史料編纂所データベース（http://wwwap.hi.u-tokyo.ac.jp/ships/db.html）

まずは、東京大学史料編纂所のデータベースを紹介します。東京大学史料編纂所は、歴史資料の編纂を古くから行っている機関です。その中で編纂の目的に即したかたちでのデータベース構築が進められており、『大日本史料』や『大日本古文書』『大日本古記録』をはじめとした、多数のデータベースを公開しています。そして、後述する奈良文化財研究所と連携することで、くずし字に関する辞書の構築を共同で実施したり、東大附属図書館と連携したデータ公開を行うなど、組織間連携も進めています。東京大学史料編纂所のデータベースは、歴史情報の発見系のデータベースとしては、ひとつの基準となっているといえるでしょう。

▶ 人間文化研究機構本部・歴史地名データ

（https://www.nihu.jp/ja/publication/source_map）

次に紹介するのは、歴博も構成機関のひとつである人間文化研究機構が出している「歴史地名データ」です。主に、明治期の地名情報の集成（大日本地名辞書・五万分の一地形図・「延喜式神名帳（えんぎしきじんみょうちょう）」）をもとにして作ったデータ群で、30 万件弱の件数を持っています。データベースとして提供されているものではなく、CSV 形式のデータをダウンロードして自由に使うことができるようになっている点が特徴です。人間文化研究機構の名前を表示すれば、それ以外は自由に利用が可能な CC BY に準拠した形式を取っています。特に明治時代の地名が緯度経度とともに表示されているため、過去の地名の位置を簡単に比定するためには有益なデータセットだといえるでしょう。

CSV でのダウンロード形式であるため、マッシュアップにも使えます。

▶ 奈良文化財研究所「木簡庫」（http://mokkanko.nabunken.go.jp/ja/）

奈良文化財研究所は「木簡庫（もっかんこ）」というデータベースを提供しています。これは、かつて木簡データベースと呼ばれていたものや、木簡字典などを統合したものになっています。木簡の情報を見つけ出すには、まずここを探す必要があります。また、近年はこの「木簡庫」のみならず、考古学の発掘調査報告書のリポジトリのデータベースを広く展開するなど、データ基盤の充実に力を注いでいます。

▶ SAT（大正新脩大蔵経テキストデータベース）（http://21dzk.l.u-tokyo.ac.jp/SAT/）

漢訳仏典の集大成として『大正新脩大蔵経（たいしょうしんしゅうだいぞうきょう）』（大正期から作成された漢訳経典の「集成」）のデータベースをあげることができます。特に、SAT では全文検索以外にも TEI・IIIF などの多様な人文情報学的展開を先駆的に行っているのが特徴であり、国際的な最新研究がどのようにデータベースの中に落ちるのか、といった実践例を見るという点でも非常に有益です。

▶ HNG（漢字字体規範史データセット）および CHISE（http://hng-data.org/）

それぞれは別物のデータベースですが、本稿執筆現在では統合的に扱われていますので、同時に紹介します。「漢字字体規範史（きはんし）データベース」（2018 年 7 月現在、長期メンテナンス中のため停止中）は、その前身である「石塚漢字字体資料」の情報カードの画像、漢字字形の切り抜き画像、およびメタデータをデータベース化したものであり、日本における漢字情報の基盤となるものでした。各時代・各地域（国）には漢字字体の標準が存在し、その標準が変遷することを実証するためのものとしても「石塚漢字字体資料」は重要です。2005 年 9 月の時点で、漢籍・仏典・国書などの典籍 67 資料、総用例数約 40 万字が収録されていました。こちらに CHISE（汎用文字符号に制約されない次世代文字処理環境の実現を目指すオープンソース型研究・開発プロジェクト）の技法が応用され、現在公開へと進められています。

そして、このプロジェクトの中で、読めない漢字を発見するために用いられるのが、漢字発見ツール「CHISE IDS Find（http://www.chise.org/ids-find）」です。ここに、読めない漢字を部首ごとに入れることで、効果的に発見し、その漢

字情報にアクセスするとともに、入力を可能にすることができます。日本における人文情報学では文字の研究も非常に盛んです。その研究の多くの事例とエッセンスを、この CHISE プロジェクトを通じて学ぶことができるでしょう。

▶ **CODH のデータセット**（http://codh.rois.ac.jp/dataset/）

情報システム研究機構の中に、人文学オープンデータセンター（CODH）という組織があります。そこからも、多くの歴史情報・人文情報に関わるデータが出されています。特に国文学研究資料館・新日本古典籍総合データベースに関わるデータを活用したものが多いです。IIIF については、その機能を十分に発揮することができる Curation Viewer を提供したり、古典籍の情報から具体的な社会展開を行った事例などもあります。

これ以外にも、東京文化財研究所や、渋沢栄一記念財団なども多くのデータを公開していますし、本章では大きすぎて触れていませんが、国立国会図書館によるデータ公開の事例などもあります。このような先行研究が、歴史情報学を推進するための重要な基準となっています。また、日本史に関わるデータベースの概観は『日本歴史』848 号（2019 年 1 月）の特集によっても知ることができます。多くの歴史研究者が作り上げたデータベースの状況を学ぶことができるでしょう。

5. おわりに

冒頭で、歴史学を正面から捉えつつ行われる歴史情報学の研究は多くはない、といいました。それ自体は事実ですが、関連する研究や基礎的なデータセットなどは多く揃いつつある状況ですので、むしろこれから研究を進めていく上では、好機であるともいえるのではないでしょうか。以降、目録・画像・テキストの基礎的な研究状況などを学ぶとともに、これらの事例を見て「このように実現されるのか」という観点を学んでもらえればと思います。